Ben Johnson Associates **DOCKET FILE COPY ORIGINAL** **August 28, 1998**

Before the
FEDERAL COMMUNICATIONS COMMISSION
Washington, D.C. 20554

	)	
	)	
Platform Development of Computer	)	CC Docket Nos. 96-45, 97-160
Models For Estimating Economic Cost	)	
	)	DA 98-1587

COMMENTS

Ben Johnson Associates, Inc. ("BJA") hereby submits its comments regarding the platform development of computer models for estimating forward-looking economic costs, as sought by the Common Carrier Bureau.

Introduction

Ben Johnson Associates, Inc. respectfully submits these comments in response to the FCC's Notice dated August 7, 1998. In our capacity as consultants to state regulators and public counsels in various states, we have performed analyses and critiques of current and prior releases of both BCPM and the HAI Model. We are thus in a position to discuss their strengths and weaknesses and respond to the various technical issues raised in the Commission's notice.

What's more, in the roughly 20 months since we submitted to the Commission an early version of our own forward-looking cost model, the Telecom Economic Cost Model, we have continued to conduct state-specific cost studies using that model, and have continued to improve and enhance the model. These experiences have taught us a great deal about the issues raised in the Notice--in particular, the problems involved in locating and clustering of customers and modeling realistic, cost-effective feeder and distribution cable routes. Since these problems are generic to all such cost modeling, and they are central to the present inquiry, we believe that a report of our experiences in attempting to solve them for ourselves may be of benefit to the Commission here. Although our model is designed to analyze network costs on a state-specific basis, with a higher level of accuracy and flexibility than is required for federal USF purposes, the problems we have faced are identical with those facing the FCC for national modeling purposes, and some of the solutions we reached may be relevant for a national application.

Customer Location Approaches

The two most important drivers of per-line network costs are average loop length and customer density. Both these drivers are functions of customer location. Therefore, if one wants to accurately identify high cost areas, and precisely measure how much higher than normal are the costs in these areas, it is crucially important to locate end-use customers with accuracy. The BCPM, Hatfield (HAI) and the Hybrid Cost Proxy (HCPM) models have all recently made attempts to incorporate geographically detailed customer location data into their algorithms. The HAI model uses clustering algorithms which process actual and surrogate geocoded data, even though important details of the location data are prematurely discarded during processing. The BCPM and HCPM gather and process data primarily on the basis of census blocks (CBs) and the HCPM appears to ignore road data in its customer location algorithms. In our opinion, all three of the models exhibit serious deficiencies in the process of locating customers. Although the models have made significant strides in the modeling of customer locations, there are still weaknesses which could be addressed with better data, better use of available road data, and the elimination of some simplifying assumptions. These weaknesses are briefly discussed below.

Customer Location Data

According to the *BCPM 3.1 Model Methodology*, the primary sources of raw data for BCPM customer location are census block (CB) data for residential customers and PNR

Associates data for business customers (mostly at the CB level). *[BCPM 3.1 Model Methodology*, p. 27.] The residential customer base used in the modeling is the census of all households, whether those households are telephone subscribers or not. This approach may introduce nonexistent economies of scale on the one hand, and build out to actually unserved areas on the other hand. Although BCPM uses all household locations to establish the required network investment, it uses actual lines to calculate the cost per line, thus creating a fundamental mismatch that inflates the latter figure.

HAI uses various data sources to estimate the number of telephone lines and their location, including geocoding of customer addresses (provided also by PNR Associates). Nevertheless, a high percentage of customers in most rural areas are not geocoded and must be captured by the "surrogate geocoding" process.

The HCPM also uses Census Block data but is moving toward the capability of processing geocoded data.[1] The HCPM developers makes no mention of how they intend to deal with non-geocodable addresses. The movement toward the use of geocoded data is a positive trend, which will ultimately provide better results–particularly as better quality data are gathered or become available.

We will now describe briefly how BJA is using geocoded customer data to locate customers. We are using an entirely data-driven, largely computerized approach. We began with telephone numbers and addresses from the white page listings. This public data source is nearly as complete, and just as reliable, as the proprietary data sources used by HAI. These data were disaggregated into residence and nonresidence (business) listings. Since detailed white page listing data are available for all wire centers, this provides a high degree of consistency. To the

[1] "In the current release, a number of small modifications have been made to 'fine tune' the model under the expectation that it will ultimately be used with a source of geocode data." *HCPM User Documentation*, p. 1.

extent feasible, we identified the exact geographic location (latitude and longitude) of each listing. The listing source data (PhoneCD) includes latitude and longitude data for many listings; this is supplemented with address matching using zip+4 and street segment data where possible. The overall accuracy and completeness of this process is comparable to that used by HAI, without the added cost and confidentiality complications involved with the proprietary data sources used by HAI. However, we would note that further improvements could be achieved, particularly regarding remote customer locations in rural areas, where, in the context of a universal service fund, a higher degree of accuracy is desired. For that purpose, we are working towards use of E911 databases, which we believe will provide substantial gains in accuracy and completeness in many rural areas. We have also gathered precise customer location data on site, using hand held GPS equipment, on a trial basis. The results of this trial effort were quite successful. We collected extremely accurate and detailed customer location data in a very low density part of the American Falls, Idaho wire center at reasonable cost.

Therefore, we make the following specific recommendations regarding customer location data.

(1) Geocoded address data from white page phone listings should be the primary means of locating customers. The census block data exclusively relied upon by BCPM and HCPM should be used as supplemental source, rather than a primary data source. The census block data is insufficiently precise and is nearly a decade old. Available "updates" of the census data are strictly estimates, which are not as accurate as the phone listing data, which is constantly updated. Alternative data sources, such as telephone carrier billing records (which contain additional addresses not listed in phone books) and E911 databases should be used to the extent feasible, since these can help reduce the geocoding failure rate.

(2) Evenly distributing ungeocoded customers along census block boundaries (HAI) is unrealistic and creates distortions; a better, less biased method of assigning locations to ungeocoded customers is needed.

(3) To the extent census data are used, adjustments should be made to distinguish between census households with phones, and those without (BCPM).

Grouping Customers

Even if the sources of customer location data could be made comprehensive, could be continually updated, and became readily available (i.e., even if we always knew exactly where the customers were), the problem would remain of how to group or cluster the customers into serving areas while meeting certain engineering constraints and criteria.

Algorithms in the BCPM model use road data to distribute customers within the wire center. The basic network geographic unit is the "ultimate grid," also called a Carrier Serving Area (CSA). After their derivation by a complex process involving "microgrids" and "macrogrids,"[2] ultimate grids are divided into four distribution quadrants, whose latitude and longitude coordinates are based on the grid's road centroid, "calculated as the average horizontal and vertical point of all roads in the defined area." (Id. p. 36, note.) Customers are apportioned to the quadrants on the basis of road mileage (a quadrant without roads gets no customers). After

[2] The entire wire center is first partitioned into "microgrids," and customers within the census block are assigned to microgrids in proportion to each microgrid's share of the CB's total area (for small CBs of less than 1/4 square mile) or in proportion to the microgrid's share of the road mileage in the census block (for CB's of 1/4 square mile area or more), excluding certain kinds of roads (limited access highways, underpasses, alleys, etc.) "where customers are unlikely to reside." [User Documentation, p. 30.] Microgrids are then aggregated into larger "macrogrids" of 64 microgrids each that establish the outer boundaries for the group. An iterative process partitions the macrogrid into four equally sized subgrids until all grids satisfy line size and technological constraints. The resulting ultimate grids have a composite household and business line count equal to the sum of the household and business lines for the associated underlying microgrids.

further processing,[3] customers in each quadrant are concentrated within a "road-reduced area" equal to1000 feet times the road distance in the quadrant. Customers are then distributed evenly within the area by dividing it into square lots according to the number of customers.

We agree that this version of BCPM is a substantial improvement over earlier versions, which distributed all customers evenly throughout a census block group (CBG). BCPM also makes a valiant effort to improve the accuracy of the locating process by tying it to the road network. However, there are at least three serious problems with the BCPM customer location process as described above: First, while the road structure provides a good starting point for distributing customers, the road system alone cannot reveal how customers are dispersed along the road. Some rural roads may have little or no population along them, yet the BCPM model assumes they have a proportionate share of customers.

Second, BCPM does not use actual customer locations, even where this information is readily available. For example, although the LECs sponsoring BCPM have internal records that include the address of many–perhaps most–of their customers, these data are completely ignored in favor of census data that are clearly inferior to more current data sources, such as white page listings or billing records.

Third, BCPM apparently distributes customers uniformly through the road-reduced area of the quadrant based on the average lot size within the quadrant. Yet, lot sizes can vary widely

[3] Within each distribution quadrant, another road centroid is established... For each non-empty distribution quadrant, the total area that falls within a 500-foot buffer along each side of the roads within that distribution quadrant is calculated. The road-reduced area is modeled as a square whose size is equal to the total road buffer area.... The center of each distribution quadrant's square road-reduced area is placed at the road centroid of the distribution quadrant.... Within each of these road-reduced areas, the customer data, apportioned at the microgrid level for housing units and business lines, is retained at the distribution quadrant level and subsequently passed to the distribution algorithm for cable design. (Id., pp. 38-9.)

even within a given area, and populations tend to cluster. BCPM appears unable to capture variations in density within census blocks, instead assuming that customers are uniformly spread through a given area in proportion to the number of road miles. This assumption clearly is invalid in, for example, the North Woods of Maine, where rural subscribers encircle Moosehead Lake and are notably absent from the many miles of logging roads. This simplification can influence the amount of distribution cable which is deployed by the model, as well as the location of feeder distribution interfaces and/or fiber electronic remote terminals. In an actual network, the latter equipment can be centrally located within population clusters, thereby minimizing costs. BCPM ignores this potential for cost savings, by assuming customers are uniformly dispersed. For instance, BCPM places remote terminals at the road centroids of ultimate grids (Id., p. 42.), rather than centrally locating them relative to the population to be served. There is no necessary relationship between the location of the road centroid and the places where customers needing to be served are the most concentrated–particularly in rural areas, where customers may be bunched along certain roads, while other roads may have very few, or any customers. If there are roads to either side of a river or lake, the road centroid, where the remote terminal is located by BCPM, may even be underwater.

Of course, road data generally provide an excellent way of estimating the amount of cable required to reach customers whose locations are known. Road data can also provide a reasonable basis for estimating the approximate location of customers that cannot be precisely located by other means (e.g., geocoding). However, by using road data in a roundabout fashion. BCPM fails to take full advantage of this potential.

In this regard, geocoding customer locations, to the extent feasible, is clearly superior. If the majority of customers are first located by geocoding of customer addresses, the specific roads and locations along roads can be identified. The remaining customers (those which can't be geocoded) can be distributed using an approach like that followed by BCPM – taking into account census data and road locations. This approach ensures that relatively few customers are mistakenly assigned to empty roads, and that the actual customer clustering patterns dominate the network modeling process (based upon geocoded data where available). However, since the BCPM does not use geocoding, this solution is not available to it. This is particularly unfortunate, since the BCPM sponsors potentially have direct access to customer billing records, 911 data bases and other data sources that are potentially superior to the public data sources used by the HAI model.

HAI 5.0a uses available geocoded address data at the earliest stages of the cost modeling process, in assigning customers to "clusters."[4] While clustering is a key step in the HAI 5.0a cost modeling process, the HAI model does not use the shape of each cluster, nor does it consider the actual location of customers within each cluster. Instead, in a step similar to the BCPM establishment of distribution areas, the HAI Distribution Module substitutes a simplified rectangular grid of equivalent size for each of the main clusters, and an algorithm assigns end-users to hypothetical rectangular lots within the overall rectangle. "The aspect ratio (height-to-width) of this rectangle is determined by the data input development process for each cluster, and distribution cable is laid out in a fashion that reflects this ratio" [HAI Model, Release 5.0a, p. 6.]

In other words, the HAI model starts with detailed customer location data, but it subsequently discards most of that information. The geocoded data are merely used to define clusters of customers; the detailed information concerning customer locations within these cluster areas is not retained. Instead, the clusters are converted into simple geometric shapes, and the customers are assumed to be evenly spaced on rectangular lots. The geocoded data are never used to locate customers within clusters, or to determine how much cable would be needed to reach all

[4] There are two types of clusters, "main" and "outlier." An outlier cluster has 1-5 lines; a main cluster has 6 lines or more. According to the HAI model documentation, In order to be considered members of a cluster, customer locations must meet the following criteria: (1) no point in a cluster may be more than 18,000 feet distant from the cluster's centroid (based on right angle routing); (2) no cluster may exceed 1,800 lines in size; and (3) no point in the cluster may be farther than two miles from its nearest neighbor in the cluster. [HAI Model, Release 5.0a, p. 32.]

of these customers (which depends upon the pattern in which they are arranged, the location and shape of rights of way, and other factors). This simplified approach potentially results in an understatement of distribution cable lengths, and thus loop costs.

Thus, although the HAI model starts with precise information concerning many customer locations within each wire center, it fails to fully utilize this information. It simplifies away or ignores important aspects of the geographic data that are potentially usable in the cost modeling process. For instance, the HAI model makes no direct use of road data, despite the fact that distribution cable generally is located along road rights of way, and it assumes customers are neatly arranged on rectangular lots, regardless of whether or not this is actually the case. The effect of these simplifications is to reduce the reliability of the HAI model in estimating distribution costs. In some cases the HAI model may overestimate and in some cases it may underestimate the amount of distribution cable required to reach customers.

Critics have also expressed concern that the clustering approach tends to understate the amount of distribution cable needed. The fundamental problem is that irregularly shaped clusters are converted into rectangles of the same area, and the customers are evenly spaced on lots within that rectangle. Under this configuration, the cable requirements can be substantially less than the cable required to serve the actual locations in the irregularly shaped clusters.

Various aspects of the HAI approach can introduce bias into the cost estimating process. For instance, since HAI assumes that all ungeocoded customers are clustered along the edges of the census block (CB), the actual amount of cable required to reach those customers will be understated wherever the customers are actually dispersed throughout the CB. The HAI approach simplistically locates ungeocoded customers on the perimeter of the CB, even if the CB is bounded by something like a river, rather than a road.

In a recent proceeding BJA participated in before the Nevada Public Service Commission, an effort was made to verify or improve the results of the HAI model by comparing the modeled loop lengths with analogous data for the actual network. In the course of this effort, substantial discrepancies were found in some wire centers. Sprint, which serves the Las Vegas area, subsequently reported on the purported inaccuracy of the HAI customer location algorithms and argued that these algorithms could result in significantly understated cabling requirements under certain conditions.

There are also simplifying assumptions and other aspects of the HAI model that can have the effect of offsetting some of these problems, and/or potentially lead to overestimates of cable lengths under some circumstances. The analysis performed in Nevada suggests that the degree and direction of error can vary depending upon the specific circumstances in each geographic area. BJA recently compared HAI feeder lengths with an inventory of actual feeder lengths of Bell Atlantic in a New Hampshire proceeding, and found very large discrepancies. In approximately half the wire centers, the HAI average feeder length differed from the Bell Atlantic actual feeder length by more than 50%, and many of these discrepancies were in the positive direction (HAI estimated more feeder length than was present in the actual network).

Although the HAI model relies upon data sources that fail to accurately geocode a sizable portion of the customer locations, this can be overcome. Every phone that is connected to the wired network has a location, and that location can potentially be identified and mapped. The geocoding "failure" rate can most easily be reduced by using additional data sources, such as the LEC's customer billing records, and/or the data base used in providing E911 service. If those data sources prove inadequate in certain areas, the next best alternative would be to gather additional data. For instance, global positioning system (GPS) satellite technology can be used to identify customer locations in sparsely populated rural areas.

The HCPM claims that, in contrast to the BCPM and HAI models, it retains the actual customer location data and avoids distortions by building distribution plant to exact customer locations, "subject to a small margin of error." [*HCPM User Documentation*, p. 1.]

The HCPM developers make an interesting point regarding the difficulty of defining the optimal size of a customer cluster:

> The objective of a clustering algorithm is to create the proper number of feasible serving areas. Unfortunately, this is not a well-defined objective, because of the existence of both fixed and variable costs associated with each additional serving area. A fixed cost gives a clear incentive to create a small number of large clusters, rather than a larger number of smaller-clusters. On the other hand, with fewer clusters the average distance of a customer from a central point of a cluster, and consequently the variable costs associated with cable and structures, will be larger. In moderate to high density areas, it is not clear, a priori, what number of clusters will embody an optimal trade-off between these fixed and variable costs. However, in low density rural areas, it is likely that fixed costs will be the most significant cost driver. Consequently, a clustering algorithm that generates the smallest number of clusters should perform well in rural areas. [HCPM User Documentation, p. 5-6.]

Granting the validity of this argument, it would appear that the HCPM is trying to build a better mousetrap (i.e. a better clustering algorithm than HAI's "nearest neighbor" algorithm).

> Clusters are evaluated on the basis of the relative distance of customers from the line weighted centroid of the new and old clusters, rather than on the basis of distance from a nearest neighbor. After an initial clustering process, two different optimization algorithms look for ways to re-assign customers to clusters, so as to reduce the total distance from the cluster centroids, while satisfying the maximum distance constraints. These optimization procedures significantly enhance the performance of the original algorithms (HCPM User Documentation, p. 1).

There are still problems, however. Since HCPM reassigns customers using airline distance as the sole criterion and the actual physical terrain features are ignored, inappropriate assignments can occur. The cluster center closest to a given customer may be on the other side of a river, mountain, or other natural obstruction. Most seriously, HCPM ignores the location and alignment of actual roads which provide feasible routes for cable. The optimal clustering taking into account roads (as the Telecom Model does) could be very different than the "optimal" clustering which ignores this important constraint.

We conclude that the following improvements in customer grouping can be made to the current versions of the BCPM, HAI and HCPM models:

(1) Retain the actual shape of clusters, rather than converting them to rectangles.

(2) Retain the actual customer locations, rather than assuming they are uniformly distributed within the rectangle/cluster.

(3) Eliminate the simplifying assumptions regarding lot sizes, if possible, or at least make them more reasonable (e.g., don't assume, as BCPM does, that lots are wider than they are deep).

(4) Incorporate road network data into the clustering and cable routing algorithms.

The current version of the Telecom Model (version 5.2) includes all of these improvements, demonstrating that they are feasible in a state-specific model. To the extent the FCC is able to also achieve these improvements in a national model, the accuracy of the results will be greatly improved.

Designing Distribution and Feeder Plant

Once the customers are accurately located and efficiently grouped into clusters the task before the modelers is to design feeder plant from the central office to the digital loop carrier (DLC) or serving area interface (SAI) and distribution plant to the customer. The HAI method of building feeder plant to its main clusters appears to be a more efficient and realistic system than the BCPM method of building feeder routes to every occupied ultimate grid, especially since some of the latter grids may contain no actual existing customers but only isolated noncustomer households. Moreover. BCPM has a tendency to deploy feeder cable along parallel routes, thereby overstating costs; a tapered "pine tree" topology is more efficient and more cost effective.

The HCPM developers spend much time describing optimization routines, Prim algorithms, "minimum distance spanning tree networks," and "pine tree networks" in coming up with the lowest cost configuration for a telecommunications network. Unfortunately, they appear to ignore the importance of various physical attributes of the territory being served, and most importantly the specific characteristics of the road network in designing the cable routes. BCPM is quite correct in recognizing that most if not all feeder and distribution plant will run along highways and streets; when this constraint is taken into account, what appears to be an "optimal" cable routing in HCPM may prove to be an impossibility in actual practice.

At the state level, BJA has been using a geographic information system (GIS) approach that combines geocoded addresses with exact road locations and other geographic data sets. A GIS is a computerized data handling and processing system which is capable of storing and using data describing places on the Earth's surface. It enables the user to analyze the spatial relationships between different data sets using location as the common attribute. User defined data layers can be combined to produce a map or perform spatial analysis functions as long as each layer is registered to a common geographic referencing system (e.g., latitude and longitude).

Central to a GIS (and distinguishing it from a computer mapping system that produces only graphic output) is a database system linking spatial data to geographic information for map features. It is based on three types of data elements: polygons, lines, and points. Behind each element is a table of attributes describing each map feature and its relationship to other features. Any item stored in the tabular relational database can be used for spatial analysis and mapping purposes in conjunction with any other map features and associated attributes. Thus a wide range of spatial and tabular information can be analyzed, stored, and updated with minimal effort and expense.

Use of GIS data for the Telecom Model focuses around three key concepts: feeder segments, nodes, and distribution areas. The model uses GIS-based data describing soil bedrock and groundwater conditions, telecom demand characteristics, and other attributes of potential distribution areas (DAs) such as area and populated road miles. Each customer within the DA is closer to that DA node, measured along existing roads, than it is to any other DA node. The DA nodes are located within population clusters, in order to minimize cable and structure costs.

The distribution areas are connected to the wire center central office using a series of feeder segments. Currently, these are generated on a schematic basis which approximates, but does not precisely follow, available rights of way.

We use the customer location data in conjunction with ARC/Info GIS software to locate each DA node and to define the geographic boundaries of each irregularly shaped distribution area. All of the customer locations within each DA have the common characteristic that they are closer to that DA node than to any other DA node, where distance is measured along available rights of way (roads). Each DA node is connected to its respective wire center by way of a series of feeder segments which are also developed using the GIS software.

Having defined the feeder routes and distribution areas, the ARC/Info GIS software is used to organize and summarize relevant data concerning each feeder segment and distribution area for inputting to the Telecom Model. These GIS data include not only spatial characteristics of the network, such as feeder segment length and distribution area size (square miles), but also the number of residential and business listings within each distribution area, and indicators of the spatial distance from these listings to the DA node via available rights of way (roads).

In short, geocoding is but the first step; once the customers are located, we use this information to determine the closest available DA node (traveling along the road network) and we also use this information to estimate distribution loop lengths and cable route lengths, based upon algorithms that assume each customer location is connected to the nearest DA node measured along the road network. This approach is far superior to the one used by HAI. Among other things, it doesn't convert irregularly shaped geographic areas into rectangles, it doesn't assume everyone lives on a rectangular lot, and it takes into account the actual distances required to reach the DA node, capturing the impact of winding roads in hilly and mountainous areas.

We make the following recommendations regarding the design of the distribution and feeder plant.

1. The distribution network should be based upon or constrained to existing roads, rather than simplified geometric assumptions.

2. Feeder/Distribution or Serving Area Interfaces should not be placed in nonsensical locations (i.e., in the middle of a river). They should be placed at or near the intersection of roads.

3. Feeder cable should be routed in an efficient, realistic (trunk and branches) topology which follows, or at least approximates, actual rights of way.

The current version of the Telecom Model (version 5.2) includes all of these improvements, demonstrating that they are feasible in a state-specific model. To the extent the FCC is able to also achieve these improvements in a national model, the accuracy of the results will be greatly improved.

Conclusions

Although none of the national models locates and groups customers as accurately as the current version of the Telecom Model, each of the models include dramatic improvements relative to previous versions. The FCC is generally heading in the right direction, particularly with regard to various improvements that have been suggested by the HCPM developers. However, none of the national models are yet capable of locating and grouping customers with acceptable accuracy, particularly in rural areas, nor do they accurately determine the amount of cable needed to connect customers to the central office. The most glaring problem with the three models currently being considered by the FCC is their failure to take full advantage of a GIS approach, particularly with regard to the actual locations of roads (as opposed to simplified data like "road centroids". In order for a model to accurately calculate the economic cost of carriers serving rural, insular and high cost areas, these problems must be addressed. If a full scale GIS approach, like that used by the Telecom Model, is too data intensive and costly for implementation at a national level, then it will at least be necessary to develop better approximations than those used in the current versions of BCPM, HAI and HCPM.

We therefore urge the Commission to concentrate its attentions on models which make better use of geocoded customer locations, preferably linked to other geographic data sets in a full scale GIS approach. Our experience with the Telecom Model has demonstrated that this approach is practical, and it yields substantial improvements in accuracy, although it is more costly and more data intensive than the simplified approaches being used by the current versions of BCPM, HAI, and HCPM.

Respectfully Submitted
BEN JOHNSON ASSOCIATES
By: ____________________
Dr. Ben Johnson
President
1234 Timberlane Rd.
Tallahassee, FL 32312